**Manuscript submission to "OMICS: A Journal of Integrative Biology"**

**Special issue on "Abiotic Stress in Plants"**

**Title**

A strategy for the identification of new abiotic stress determinants in Arabidopsis using web-based data mining and reverse genetics

**Authors**

Herlânder Azevedo[1,§,*] , Joana Silva-Correia[1,§,†] , Juliana Oliveira[1] , Sara Laranjeira[1] , Cátia Barbeta[1], Vitor Amorim-Silva[1] , Miguel A. Botella[2] , Teresa Lino-Neto[1] , Rui M. Tavares[1]

**Affiliation**

[1] Center for Biodiversity, Functional & Integrative Genomics (BioFIG), CBFP/Department of Biology, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

[2] Laboratorio de Bioquímica y Biotecnología Vegetal Edificio I+D 3 Planta Departamento Biología Molecular y Bioquímica Universidad de Málaga, 29071 Málaga, Spain

[§] Both authors contributed equally to the work

[†] Current address: 3B's Research Group – Biomaterials, Biodegradables and Biomimetics, University of Minho, Headquarters of the European Institute of Excellence on Tissue Engineering and Regenerative Medicine, AvePark, 4806-909 Taipas, Guimarães, Portugal

* Corresponding author

**Keywords**

*Arabidopsis thaliana*; Functional discovery; Heat stress; Reverse genetics; Web-based resources

**Full contact information of all authors**

Herlânder Azevedo

Email: hazevedo@bio.uminho.pt

Tlf: +351 253 601547

Fax: +351 253 678980

Joana Silva-Correia

Email: joana.correia@dep.uminho.pt

Tlf: +351 253 510911

Fax: +351 253 510909

Juliana Oliveira

Email: jules.olive@bio.uminho.pt

Tlf: +351 253 601544

Fax: +351 253 678980

Sara Laranjeira

Email: saranjeira@bio.uminho.pt

Tlf: +351. 253 604047

Fax: +351 253 678980

Cátia Barbeta

E-mail: catiabarbeta@bio.uminho.pt

Tlf: +351 253 601544

Fax: +351 253 678980

Vitor Amorim-Silva

Email: vitor.amorim.silva@bio.uminho.pt

Tlf: +351 253 604047

Fax: +351 253 678980

Miguel A. Botella

Email: mabotella@uma.es

Tlf: +34 952 131938

Fax: +34 952 134267

Teresa Lino-Neto

Email: tlneto@bio.uminho.pt

Tlf: +351 253 601544

Fax: +351 253 678980

Rui M. Tavares

Email: tavares@bio.uminho.pt

Tlf: +351 253 601544

Fax: +351 253 678980

**Abstract**

Since the sequencing of the *Arabidopsis thaliana* genome in 2000, plant researchers have faced the complex challenge of assigning function to thousands of genes. Functional discovery by *in silico* prediction or homology search resolved a significant number of genes, but only a minor part has been experimentally validated. Arabidopsis entry into the post-genomic era signified a massive increase in high-throughput approaches to functional discovery, which have since become available through publicly-available web-based resources. The present work focuses on an easy and straightforward strategy that couples data-mining to reverse genetics principles, to allow for the identification of new abiotic stress determinant genes. The strategy explores systematic microarray-based transcriptomics experiments, involving Arabidopsis abiotic stress responses. An overview of the most significant resources and databases for functional discovery in Arabidopsis is presented. The successful application of the outlined strategy is illustrated by the identification of a new abiotic stress determinant gene, *HRR*, which displays a heat stress-related phenotype after a loss-of-function reverse genetics approach.

## Introduction

In the wake of the international coordinated effort to sequence the first genome of a higher plant (*Arabidopsis thaliana*) (Arabidopsis Genome Initiative, 2000), plant research entered a stimulating era and gained a new leading focus: assigning a biological function to thousands of annotated genes. In this, the model plant Arabidopsis displays powerful attributes for gene function assessment: a fast and compacted growth, small genome, simplicity of genetic manipulation, together with the efficient ability to be transformed by Agrobacterium. Moreover, a variety of new resources for functional discovery (databases, analysis tools, cDNA, seed and mutant line collections) soon became available as a result of a well-organized scientific community (Feng and Mundy, 2006).

Phenotype-centered approaches have become a standard in functional discovery, namely through forward and reverse genetic strategies (Alonso and Ecker, 2006). Forward genetics (in which the mutant population is screened for a phenotype of interest) was the earlier and most convenient strategy. Though it led to the main functional discoveries up to this point, it is approaching saturation of the possible screenable phenotypes. Reverse genetics, which goes from gene selection to detection of a visible phenotype, is currently the most widespread methodology. This evolution is due to the massive use of insertion mutants using biological vectors (mainly T-DNA or transposons). In the presence of a sequenced genome, these vectors combine a low insertion-per-plant ratio to the speedy mapping of the mutation (Feng and Mundy, 2006). In Arabidopsis, large collections of gene indexed insertion mutants are now available. These systematic and for the most part public collections are almost at a genome-saturation point. More precisely, in 2010 almost 96% of Arabidopsis unique genes (27543 of a total of 28691) comprised at least one sequenced insertion element,

and 62% (17721 of a total of 28691) presented one or more confirmed homozygous insertions (Masc Report, 2010). Furthermore, 95% of the genes (27257 of a total of 28691) present confirmed expression, through different transcriptomics approaches, including small RNA sequencing projects (Masc Report, 2010). As a consequence, a loss-of function analysis of a gene-of-interest (GOI) is now fairly straightforward.

Abiotic stress has been the focus of intense research, mainly because of current climate changes that endanger worldwide agricultural yield production and produces annual losses of billions of dollars. The main abiotic stresses affecting plants have been thoroughly studied in Arabidopsis, including drought, salinity, heat, cold, chilling, high light intensity, all with a main common element that is water availability (Mittler and Blumwald, 2010). Nevertheless, knowledge on the capacity of plants to cope with all these stresses is still clearly insufficient, as many stress-related genes are still functionless in the whole-plant concept (Ahuja et al., 2010). Research in Arabidopsis, with its high-throughput and omics-based approaches, has been key to understanding molecular processes and networks involved in stress tolerance, quickly translating this knowledge to other plant species (Century et al., 2008; Masc Report, 2010).

In recent years, microarray transcriptomics technology has become a standard method to analyse gene expression at the whole-genome level. The successful use of gene expression arrays has allowed for the detection of qualitative differences resulting from the exposure to various stimuli in different plant species (Wullschleger and Difazio, 2003). In Arabidopsis, the study of stress responses at a genomic level has been greatly improved by expression profile analysis, from which several genes playing a role in wounding, cold, salt, drought and heat stresses have been identified (Cheong et al., 2002; Rizhsky et al., 2004; Oono et al., 2006; reviewed by Sreenivasulu et al., 2007; Swindell et al., 2007). In the post-genomic era, a huge amount of *in silico* data has

become available in Arabidopsis, including systematic microarray analysis of fundamental plant processes or the profiling of various knockout mutants (Kilian et al., 2007). The foremost example is the *AtGenExpress* project, a systematic transcriptomics study in Arabidopsis conducted using the Affymetrix ATH1 microarray chip. ATH1 allows transcript profiling of ~24,000 Arabidopsis genes using the Affymetrix one-colour microarray gene expression technology (Redman et al., 2004). The different experimental approaches deposited in this project were not problem-oriented or focused on specific sets of genes, thus providing high fidelity and comparable data with detailed quantitative and physiological kinetics analyses. These datasets can be further complemented with other systematic transcriptomics analysis, including the development expression map and the response to plant pathogens, among others (Kilian et al., 2007; Goda et al., 2008). Data mining of this information, easily acquired from publicly available databases, can now allow the identification of genes whose functions are putatively linked to a plant process of interest. The putative determinants frequently lack functional characterization, which makes them attractive GOIs for reverse genetics studies.

While access to a GOI's mutant line is now an undemanding step in a reverse genetics approach, the search for a phenotype is not quite as simple. In the present work we detail a straightforward data-mining strategy that can be translated into the identification of an abiotic stress-related phenotype in a gain- or loss-of-function mutant of a previously unresolved gene. This transcriptomics-based strategy is particularly adequate for unbiased researchers that are set to initiate the functional characterization of novel genes and could benefit from the recently acquired convenience of reverse genetics in Arabidopsis. Given the profusion of publicly available microarray data, the strategy can be adapted beyond abiotic stress to virtually all aspects of plant physiology.

A small overview of fundamental web-based resources supporting the present strategy is also outlined. The identification of *HRR* and detection of a heat stress-responsive phenotype in the *hrr* loss-of-function mutant validates the proposed strategy.

## Materials and Methods

### *Bioinformatic analysis*

Raw microarray data, in the form of spreadsheet-based files, was downloaded from NASCArrays (affymetrix.arabidopsis.info/narrays/experimentbrowse.pl). Spreadsheet software (Excel) was used for data mining, transformation and sorting. The *heat stress time course* experiment of the *AtGenExpress: Abiotic stress series* (Kilian et al., 2007) was carried out in 16-day-old *Arabidopsis thaliana* seedlings, with heat stress treatment starting within 3 hours of the light period. The experiment was as follows: heat stress at 38°C, samples taken at 0.25, 0.5, 1.0, 3.0 h; recovery at 25°C, samples taken at +1, +3, +9, +21 h. Databases used for web-based data mining and functional information recovery are described in high detail elsewhere (Table 1).

Venn analysis was carried out using Venn Diagram Generator (www.pangloss.com/seidel/Protocols/venn.cgi). Hierarchical clustering analysis (Eisen et al., 1998) was performed using an Euclidean distance metric, on gene expression ratios during the time course of heat stress imposition in comparison to the corresponding control. For this purpose the MultiExperiment Viewer v4.0 of the TM4 software suit was used (Saeed et al., 2003). Clustering Analysis of gene expression patterns in response to different abiotic stresses and Meta-Profile Analysis of gene expression in different Arabidopsis tissues was carried out using Genevestigator (Hruz et al. 2008). GO categorization was carried out at TAIR (www.arabidopsis.org). Prediction of sub-cellular localization used the Cell eFP Browser functionality of BAR (Winter et al. 2007).

*Plant material and growth conditions*

The *HRR* mutant line (*hrr*), a transposon line (ref. GT_5_47364) generated in *Arabidopsis thaliana Landsberg erecta* (*Ler*) background by the John Innes Centre (JIC, UK), was used to evaluate the effect of *HRR* (At5g53680) loss-of-function. Both the *hrr* and *Ler* lines were ordered through the NASC European Arabidopsis Stock Centre (arabidopsis.info/). Plants were regularly grown in a 4:1 soil:vermiculite mixture, at 23ºC and a 16 h light/ 8 h dark photoperiod (80 µE $m^{-2}$ $s^{-1}$ light intensity). The *hrr* mutant was genotyped by diagnostic PCR following JIC recommendations.

*HRR overexpression line JP5*

The *JP5* overexpression line was obtained by transformation of the *hrr* mutant with a *p35S::HRR.1:GFP6* construct. The construct followed a Gateway-based cloning strategy (Invitrogen). For this purpose, RNA was extracted from 16-day-old *Ler* seedlings subjected to a 38ºC heat shock for one hour, using the *TRIzol reagent* (Invitrogen) according to the manufacturer's instruction. The cDNA was synthesized (*SuperScript First-strand Synthesis System*; Invitrogen) and subsequently used in an RT-PCR reaction to amplify the HRR.1 coding sequence. Primers included the *attb* Gateway recombination region (highlighted in bold) and were as follows: *attB1_HRR.1*, 5'-**GGGGACAAGTTTGTACAAAAAAGCAGGCT**TAGACATGTCTCACCACCACC AAAAC-3'; *attB2_HRR.1*, 5'-**GGGGACCACTTTGTACAAGAAAGCTGGGT**AGCGAAGATCCCGGTGTCGA AAG-3'. The PCR product (HRR.1 cDNA flanked by *attb* regions) was sequentially cloned by BP recombination reaction into a pDONR 201 entry vector (Invitrogen), and

LR recombination reaction into a pMDC83 destination vector (Curtis and Grossniklaus, 2003), thus generating a HRR.1:GFP fusion under regulation of the strong constitutive 35S promoter. *Agrobacterium tumefaciens* strain EHA105 was used in the transformation of the *hrr* background by the floral dip method (Clough and Bent, 1998). A resistance marker (Hygromycin) strategy was employed to select for homozygous transformants in the T3 generation. Overexpression strength of several independent lines was evaluated by semi-quantitative RT-PCR using internal HRR.1 primers, as well as the previously described RNA/cDNA procedures, allowing for the selection of the *JP5* overexpression line.

*Heat tolerance germination assay*

Synchronised *Ler*, *hrr* and *JP5* plants were stratified at 4ºC for 3 days and subsequently surface sterilised (Weigel and Glazebrook, 2002). For the heat shock (HS) treatment, sterilised seeds were incubated at 50ºC for 60 min. Immediately after HS, seeds were ressuspended in sterile 0.25% (w/v) agarose solution and sown onto agarised Murashige and Skoog (MS) medium (Murashige and Skoog, 1962). The plates were incubated under a 16 h light/ 8 h dark photoperiod (80 µE $m^{-2}$ $s^{-1}$ light intensity) at 23ºC. The emergence of radicle was evaluated daily between days 2-10 after HS. The germination rate (%) was normalised with the corresponding germinated seeds under control conditions (25ºC). Mean and SEM were determined based on results from four independent replicates for each seed line, all containing 30 seeds. Experiments were repeated with similar results. Results were subjected to statistical analysis using a one-way-ANOVA (95% confidence interval, with Tukey Post test for column comparison; *GraphPad Prism* v5.0 software).

## Results and Discussion

*Differential expression analysis of public microarray data*

With the onset of omics-based strategies, an almost overwhelming amount of *in silico* information became available to help functional discovery in Arabidopsis as well as other plant species. Insights into biological problems-of-interest are now easily gained through web-based resources. Reviews of these genomic, epigenomic, transcriptomic, proteomic, and metabolomic data sets have been elsewhere performed (Lu and Last, 2008; Brady and Provart, 2009). However, an overview and brief description of databases essential to the present strategy is included in this report (Table 1).

Transcriptomics data from the *AtGenExpress: Abiotic stress series* consists of the global spatial-temporal gene expression pattern of the Arabidopsis response to stresses such as heat, cold, drought, salt, osmotic, wounding and UV-B light stress (Kilian et al., 2007). Web-based tools such as Genevestigator and BAR (Table 1) allow this data to be explored using user-friendly interfaces. However, they are designed for a gene-centered approach. Therefore, in order to screen for new abiotic stress determinants, namely for previously unresolved genes, the present strategy made use of the downloadable raw microarray datasets, available at various resources (Table 1). More precisely, the complete and time-resolved transcriptome data of various abiotic stresses within the *AtGenExpress: Abiotic stress series* was accessed through NASCArrays. Present results focus on the analysis of the *heat stress series*, performed on seedling shoots and roots. Data from a similar experimental design performed on suspension cells was also retrieved.

An initial search premise was defined: “Identification of novel, functionally unresolved genes that are constitutively and specifically involved in the heat stress

response, operating at the regulatory/cell nuclear level". An outline of the entire strategy is depicted in Figure 1. A straightforward raw data analysis was adopted, based on the identification of genes that (1) were differentially expressed and (2) possessed significant expression levels. Analysis was performed on standard spreadsheet-based software (*e.g.* Excel). Given the one-colour nature of the ATH1 microarray, absolute expression values for each gene/probe set were represented in the form of pixel count. Differentially expressed genes were selected based on a two-fold cut-off value in the expression ratios between stress and control experiments. Subsequently, genes were isolated that presented >500 pixel count for stress-related datasets (in up-regulated genes) and control datasets (in down-regulated genes), as previously described (Goda et al., 2002; Moseyko et al., 2002). This analysis was performed for each sampling time.

Given the primary interest on heat responsive genes, a time-course analysis of differentially up-regulated genes was performed (Fig. 2). Considering the high number of genes with two-fold enhanced expression in heat stressed cell suspensions, a ten-fold relative expression threshold was subsequently defined. As depicted in Figure 2, the present approach is sufficient to single out a great number of genes with strong putative involvement in the biological process of interest, while avoiding extensive data analysis. In heat stressed plants and cell suspensions, the number of up-regulated genes rapidly increased after the stress onset and continued to rise until the end of stress imposition (3 h). During recovery at 25°C, the number of genes displaying relative expression values above two rapidly dropped off. Given that the experimental procedure of independent microarray experiments will often produce different ranges of signal intensity (pixel count), the expression level/pixel count cut-off value should be adjusted to promote a significant number of genes for further analysis while removing low-expression genes. The cut-off value of 500 pixel count proved effective for data analysis

from five other abiotic stresses (data not shown), in which it was helped by the systematic nature of the *AtGenExpress* experimental setup. Performing the absolute gene expression cut-off is of importance, since lack of sensitivity in the signal/pixel count of low-expressing genes will often produce unrealistically high differential expression ratios that bias the analysis. The downside may be the exclusion of potentially important genes, but the high number of unresolved genes that are still identified guarantee the viability of the strategy.

Based on the initial premise, genes were single out that evidenced up-regulation during the 3h heat stress period in all tissues (shoots, roots and suspension cells). After discarding genes belonging to mitochondrial or plastidial genomes (ten genes), analysis resulted in a total of 823 genes. At this stage, the followed strategy still displayed a great number of genes. Adjusting cut-off values could have helped to reduce this number, but a second stage of more focused and conscious analysis of the data should be more appropriate to narrow down to a limited number of GOI with great potential. In order to identify genes that were strongly involved in the heat stress response, we cross-referenced genes from all three tissues to single out common up-regulated genes, resulting in 137 genes over-expressed simultaneously in all tissues (Fig. 2B).

*Identification of new heat stress determinants*

Further analysis is helped by the existence of user-friendly web-based software that greatly facilitates the interface with the information. Multiple gene analysis is possible with hierarchical clustering of genes, automatically evidencing gene clusters displaying the behavior of interest. Two main resources with these functionalities are presently available, “Clustering Analysis” within Genevestigator and “Expression Browser”

within BAR (Table 1). Hierarchical clustering analysis was performed on the time-course transcript profile of the 137 selected genes (Fig. 3), allowing these to be grouped according to similar expression behaviors, and thus differentiating classes of genes (Eisen et al., 1998). A similar response profile was found in shoots and roots. Transcription in suspension cells clearly responded more quickly to heat. This is likely due to the experimental setup allowing for a higher capacity of these cells to sense a shift in temperature, together with their homogenous nature. The fact that a ten-fold cut-off in relative expression values was required to singled out the most heat-responsive genes (Fig. 2) also reflects this behaviour. The overall response was naturally biased by the fact that the genes were previously selected for evidencing strong expression during the 3h heat stress period in all tissues. None withstanding, results clearly separate between 3h heat-shock-specific transcripts and long term (12h) responsive genes. After 21h of recovery, most genes returned to regular expression values, though suspension cells evidenced a significantly higher number of long term expression genes.

The physiological response of plants to heat stress depends on the initial perception of temperature change, triggering signal transduction pathways that lead to transcriptional events, and ultimately to cellular homeostasis recovery (reviewed by Wahid et al., 2007). Under heat stress, two different groups of genes have been identified: those responsible for regulation of gene expression and signal transduction pathways, and those directly associated with protection against stress (reviewed by Kaur and Gupta, 2005).The first group includes transcription factors, protein kinases and phosphoinositide metabolism-associated enzymes. These are usually early response genes, induced quickly and transiently to activate the delayed response genes. The second group of effector proteins can include osmoprotectant enzymes, antifreeze proteins, chaperones and scavenging enzymes (reviewed by Kaur and Gupta, 2005).

Given the initial premise, hierarchical clustering allowed us to confirm most genes as being early responsive (expressed within one hour of heat imposition) and are likely to be part of transcriptional regulatory networks.

In Arabidopsis, the existence of systematic transcriptomics data (*e.g.* development map, hormone response or biotic stress) provides an excellent coverage of the main aspects of plant physiology. The list of putative GOIs can thus be reduced by conducting expression pattern analysis regarding additional specificities. The expression pattern of 137 genes was analyzed in response to diverse abiotic stress stimuli, in order to single out heat-specific transcripts in comparison to additional environmental stress factors. Using Genevestigator, an electronic Northern of these genes under several abiotic stresses (cold, osmotic, salt, drought, genotoxic, oxidative, UV-B, wounding and heat) was performed (Fig. 4). Hierarchical clustering (Pearson's correlation coefficient) of the transcript profiles was included in the analysis. Gene clusters that presented a higher heat stress-specific signature were singled out, allowing for the identification of 43 genes as putative determinants for heat-specific response in Arabidopsis. A detailed list of selected genes is enclosed in the Supplementary Table S1.

The conjunction of sequenced genomes, omics-based approaches and *in silico* prediction allows current researchers to access a wealth of functional data, even before reaching the wet-lab. As previously stated, several reviews provide excellent coverage of this aspect (Lu and Last, 2008; Brady and Provart, 2009). Therefore, the present report will focus solely on aspects of functional categorization that were used in the final stage of GOI selection. Genevestigator (Meta-Profile Analysis - Anatomy) was used to map expression of the genes across the full range of Arabidopsis tissues (Fig. 5A). With few exceptions, the selected genes displayed low expression on different tissues. As heat-stress specific genes, it is expectable that under standard conditions

expression levels are indeed reduced. The 43 heat-stress specific genes were then analyzed according to their functional classification using *Gene Ontology* (Ashburner et al., 2000) (Fig. 5B). The predicted assignments for each selected gene were provided by TAIR, but a similar functionality is easily available through various resources. Functional annotations are described using the specific GO terms, according to the biological process, cellular component and molecular function. Despite the majority of selected heat-stress specific genes not having predicted functional information, a broad range of functional categories were still observed (Fig. 5B). According to the biological process, many of the selected genes appear to be strongly implicated in the stress response, some of which have already been associated to the heat stress response (Supplementary Table S1). Additionally, various genes have a predicted role in protein metabolism (mainly in protein folding), which is widely known to be an important adaptive mechanism to heat stress (reviewed by Wahid et al., 2007). More importantly, the majority of the genes do not seem to be associated to a known biological process or a cellular component (Fig 5B). Concerning the subcellular localization, most genes seem to be chloroplast-targeted, intracellular or associated to membranes. Remaining genes are included in diverse cellular component categories (nucleus, mitochondrion, apoplast and cell wall). Regarding the molecular function, the majority of the selected genes display binding capacity to a wide variety of molecules (nucleotide, protein, zinc ion, RNA, auxin, aminoacid and metal clusters). Additionally, a few number of genes present catalytic activity.

For determining unresolved heat-specific genes, the state-of-the-art for each heat-stress associated gene was evaluated by performing a literature search in TAIR and PubMed. A search was directed towards genes with low/inexistent functional knowledge, or whose implication in the heat stress response had not been proposed up

to that point. By combining the gathered functional information, the GOI list was narrowed down to 31 functionally unresolved genes (Supplementary Table S1). Finally, the predicted sub-cellular targeting was estimated for each gene using the Cell eFP Browser tool (BAR; Table 1). Based on the initial premise of identifying genes with putative nuclear localization or involvement in nuclear-related processes, the GOI list was finally reduced to 20 genes.

*Functional characterization of HRR, a putative heat-responsive gene*

Taking advantage of the powerful reverse genetics resources available in Arabidopsis (Feng and Mundy, 2006), identification of phenotypes-of-interest was initiated for selected genes using loss-of-function mutants. An overview of the main resources involved in identifying and ordering loss-of-function insertion mutants is included in Table 1. Viability of this strategy can be exemplified by the identification of the functionally unresolved *HRR* (*Heat-Responsive RNA Recognition Motif*) gene. *HRR* (At5g53680) codes for a protein containing a RNA recognition motif (RRM) and up to this moment no relevant publications are known. Using the originally accessed microarray data, the expression response profile was analysed (Fig. 6A). Induction of gene expression was observed just after heat stress imposition in roots and cell suspensions, where maximal expression levels were detected as soon as one hour after heat stress imposition. Expression levels were significantly reduced during recovery at 25°C. Analysis using the Cell eFP Browser (BAR), a resource that compiles various sub-cellular targeting prediction algorithms, was performed for *HRR* (Fig.6B). Prediction of targeting to the nucleus was observed, although with a fairly low confidence score. Prediction to the cytosol is also observable, although with minimum confidence. Additional functional data includes a co-expression gene network enriched

in genes implicated in transcriptional regulation and heat stress response, as well as a GO categorization as an unknown protein (GO Biological process; GO Cellular component) with RNA binding capacity (GO Molecular function) (data not shown).

The T-DNA Express tool was the starting point of a reverse genetics strategy to characterize *HRR*, by allowing the search for insertion mutants within this genomic regions (see Table 1 for additional resources-of-interest). As a result, a loss-of-function insertion mutant was found, interrupting this gene in the third exon (Fig. 6C). The mutant was subsequently screened for heat stress-related phenotypes compared to the wild-type *Ler* background, revealing a significantly lower *in vitro* seed germination rate when seeds were subjected to a heat shock (50ºC, 60 min) (Fig. 6D). Results implicate HRR in the protective response of seeds against heat stress. Confirmation of this genotype-phenotype association was performed by transforming the mutant with an overexpressing copy of *HRR*, which resulted in the restoration of the wild-type phenotype (Fig. 6D).

## Conclusions

As climatic changes increase average global temperatures, the study of plant abiotic stress and in particular of heat responsive mechanisms is of major importance. One of the immediate focuses of plant research has been the understanding of the highly complex abiotic stress-responsive networks and the functional characterization of associated genes. The vast array of Arabidopsis-based resources and tools that can now be easily accessed using the internet offers the researcher a powerful amount of functional data, helping him in the task of gene function discovery. How this data ultimately translates into a gain of knowledge on plant physiology and molecular biology remains a challenge. Using publicly available transcriptomics data, the proposed strategy relies on the identification of differentially expressed genes under abiotic stress imposing conditions, followed by data-mining of functional information to identify potentially new determinants. The strategy was exemplified using heat stress, but it has been extended to various abiotic stresses, being currently employed in our laboratory for hypothesis generation and functional discovery of new abiotic stress determinants.

The efficiency of AtGenExpress datasets for the identification of heat shock proteins and transcription factors, specifically involved in the heat response pathway, has already been reported (Swindell et al., 2007). A significant number of differentially expressed transcripts detected after heat treatment have also provided helpful insights into plant heat stress responses (reviewed by Huang and Xu, 2008). In the present work, a straightforward and accessible microarray data analysis led to the identification of a great number of heat stress up-regulated transcripts. For the selection of more specific determinants, several tools were applied, including expression pattern analysis and

cross-referencing of functional information. Ultimately, this led to the identification of a previously unresolved gene, *HRR*, which was shown to be determinant for seed viability after heat shock treatment. Functional characterization of *HRR* is now underway.

The present strategy can be adapted to complement ongoing research efforts or suit unbiased researchers keen on novel functional discovery, particularly in those aspects involving plant abiotic stress responses.

**Acknowledgments**

No competing financial interests exist. The present work was supported by *Foundation for Science and Technology* (POCTI/AGR/45462/2002). H. Azevedo (SFRH/BPD/17198/2004), J. Correia (SFRH/BD/16663/2004), J. Oliveira (SFRH/BD/38379/2007), S. Laranjeira (SFRH/BD/29778/2006), C. Barbeta (SFRH/BD/12081/2003) and V. Amorim-Silva (SFRH/BD/29778/2006) were supported by *Foundation for Science and Technology*.

**Author Disclosure Statement**

No competing financial interests exist.

**References**


AHUJA, I., VOS, R. C. H. D., BONES, A. M., AND HALL, R. D. (2010). Plant molecular stress responses face climate change. Trends Plant Sci **15**, 664-674.

ALONSO, J. M., AND ECKER, J. R. (2006). Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in *Arabidopsis*. Nature Rev Genet **7**, 524-536.

ARABIDOPSIS GENOME INITIATIVE (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**, 796-815.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., ET AL. (2000). Gene Ontology: tool for the unification of biology. Nature Genet **25**, 25-29.

BARRETT, T., TROUP, D. B., WILHITE, S. E., LEDOUX, P., EVANGELISTA, C., KIM, I. F., ET AL. (2011). NCBI GEO: archive for functional genomics data sets-10 years on. Nucleic Acids Res **39**, D1005–D1010.

BRADY, S. M., AND PROVART, N. J. (2009). Web-Queryable Large-Scale Data Sets for Hypothesis Generation in Plant Biology. Plant Cell **21**, 1034-1051.

BRAZMA, A., HINGAMP, P., QUACKENBUSH, J., SHERLOCK, G., SPELLMAN, P., STOECKERT, C., ET AL. (2001). Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. Nature Genet **29**, 365-371.

CENTURY, K., REUBER, T. L., AND RATCLIFFE, O. J. (2008). Regulating the Regulators: The Future Prospects for Transcription-Factor-Based Agricultural Biotechnology Products. Plant Physiol **147**, 20–29.

CHEONG, Y. H., CHANG, H. S., GUPTA, R., WANG, X., ZHU, T., AND LUAN, S. (2002). Transcriptional profiling reveals novel interations between wounding, pathogen, abiotic stress, and hormonal responses in Arabidopsis. Plant Physiol **129**, 661-677.

CLOUGH, S. J., AND BENT, A. F. (1998). Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. Plant J **16**, 735–743.

CRAIGON, D. J., JAMES, N., OKYERE, J., HIGGINS, J., JOTHAM, J., AND MAY, S. (2004). NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res **32**, D575-D577.

CURTIS, M.D., AND Grossniklaus, U. (2003). A Gateway cloning vector set for high-throughput functional analysis of genes in planta. Plant Physiol **133**, 462–469.

DUVICK, J., FU, A., MUPPIRALA, U., SABHARWAL, M., WILKERSON, M. D., LAWRENCE, C. J., ET AL. (2008). PlantGDB: a resource for comparative plant genomics. Nucleic Acids Res **36**, D959–D965.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O., AND BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences **95**, 14863–14868.

FENG, C. P., AND MUNDY, J. (2006). Gene discovery and functional analyses in the model plant *Arabidopsis*. JIPB **48**, 5-14.

GODA, H., SASAKI, E., AKIYAMA, K., MARUYAMA-NAKASHITA, A., NAKABAYASHI, K., LI, W., ET AL. (2008). The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. Plant J **55**, 526–542.

GODA, H., SHIMADA, Y., ASAMI, T., FUJIOKA, S., AND YOSHIDA, S. (2002). Microarray Analysis of Brassinosteroid-Regulated Genes in Arabidopsis. Plant Physiol **130**, 1319–1334.

HILSON, P., ALLEMEERSCH, J., ALTMANN, T., AUBOURG, S., AVON, A., BEYNON, J., ET AL. (2004). Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications. Genome Res **14**, 2176-2189.

HRUZ, T., LAULE, O., SZABO, G., WESSENDORP, F., BLEULER, S., OERTLE, L., ET AL. (2008). Genevestigator V3: A Reference Expression Database for the Meta-Analysis of Transcriptomes. Adv Bioinformatics **2008**, 1-5.

HUANG, B., AND XU, C. (2008). Identification and characterization of proteins associated with plant tolerance to heat stress. JIPB **50**, 1230–1237.

KAUR, N., AND GUPTA, A. K. (2005). Signal transduction pathways under abiotic stresses in plants. Curr Sci **88**, 1771-1780.

KERSEY, P. J., LAWSON, D., BIRNEY, E., DERWENT, P. S., HAIMEL, M., HERRERO, J., ET AL. (2010). Ensembl Genomes: Extending Ensembl across the taxonomic space. Nucleic Acids Res **38**, D563–D569.

KILIAN, J., WHITEHEAD, D., HORAK, J., WANKE, D., WEINL, S., BATISTIC, O., ET AL. (2007). The AtGenExpress global stress expression data set: protocols, evaluation and model analysis of UV-B light, drought and cold stress responses. Plant J **50**, 347-363.

LI, Y., ROSSO, M. G., VIEHOEVER, P., AND WEISSHAAR, B. (2007). GABI-Kat SimpleSearch: an *Arabidopsis thaliana* T-DNA mutant database with detailed information for confirmed insertions. Nucleic Acids Res **35**, D874–D878.

LU, Y., AND LAST, R. L. (2008). Web-Based Arabidopsis Functional and Structural Genomics Resources. *The Arabidopsis book*. e0118 doi: 10.1199/tab.0118.

MASC REPORT (2010). Annual Report The Multinational Coordinated *Arabidopsis thaliana* Functional Genomics Project

MEWES, H. W., RUEPP, A., THEIS, F., RATTEI, T., WALTER, M., FRISHMAN, D., ET AL. (2010). MIPS: curated databases and comprehensive secondary data resources in 2010. Nucleic Acids Res **Advanced Acess**, 1-5.

MITTLER, R., AND BLUMWALD, E. (2010). Genetic Engineering for Modern Agriculture: Challenges and Perspectives. Annu Rev Plant Biol **61**, 443–462.

MOSEYKO, N., ZHU, T., CHANG, H. S., WANG, X., AND FELDMAN, L. J. (2002). Transcription profiling of the early gravitropic response in Arabidopsis using high-density oligonucleotide probe microarrays. Plant Physiol **130**, 720-728.

MURASHIGE, T., AND SKOOG, F. (1962). A revised medium for rapid growth and bioassay with tobacco issue cultures. Physiol Plant **15**, 473-497.

OONO, Y., SEKI, M., SATOU, M., IIDA, K., AKIYAMA, K., SAKURAI, T., ET AL. (2006). Monitoring expression profiles of Arabidopsis genes during cold acclimation and deacclimation using DNA microarrays. Funct Integr Genomics **6**, 212-234.

PROVART, N., AND ZHU, T. (2003). A Browser-based Functional Classification SuperViewer for Arabidopsis Genomics. Currents in Computational Molecular Biology **2003**, 271-272.

REDMAN, J. C., HAAS, B. J., TANIMOTO, G., AND TOWN, C. D. (2004). Development and evaluation of an *Arabidopsis* whole genome Affymetrix probe array. Plant J **38-561**, 545.

RHEE, S. Y., BEAVIS, W., BERARDINI, T. Z., CHEN, G., DIXON, D., DOYLE, A., ET AL. (2003). The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community Nucleic Acids Res **31**, 224-228.

RIZHSKY, L., LIANG, H., SHUMAN, J., SHULAEV, V., DAVLETOVA, S., AND MITTLER, R. (2004). When defense pathways collide. The response of *Arabidopsis* to a combination of drought and heat stress. Plant Physiol **134**, 1683-1696.

SAEED, A. I., SHAROV, V., WHITE, J., LI, J., LIANG, W., BHAGABATI, N., ET AL. (2003). TM4: A Free, Open-Source System for Microarray Data Management and Analysis. BioTechniques **34**, 374-378.

SAMSON, F., BRUNAUD, V., BALZERGUE, S., DUBREUCQ, B., LEPINIEC, L., PELLETIER, G., ET AL. (2002). FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of *Arabidopsis thaliana* T-DNA transformants. Nucleic Acids Res **30**, 94-97.

SCHMID, M., DAVISON, T. S., HENZ, S. R., PAPE, U. J., DEMAR, M., VINGRON, M., ET AL. (2005). A gene expression map of *Arabidopsis thaliana* development. Nature Genet **37**, 501 - 506.

SCHOLL, R. L., MAY, S. T., AND WARE, D. H. (2000). Seed and molecular resources for Arabidopsis. Plant Physiol **124**, 1477-1480.

SCHWAB, R., OSSOWSKI, S., RIESTER, M., WARTHMANN, N., AND WEIGEL, D. (2006). Highly Specific Gene Silencing by Artificial MicroRNAs in Arabidopsis. Plant Cell **18**, 1121–1133.

SREENIVASULU, N., SOPORY, S. K., AND KISHOR, P. B. K. (2007). Deciphering the regulatory mechanisms of abiotic stress tolerance in plants by genomic approaches. Gene **388**, 1-13

SWARBRECK, D., WILKS, C., LAMESCH, P., BERARDINI, T. Z., GARCIA-HERNANDEZ, M., FOERSTER, H., ET AL. (2008). The Arabidopsis Information

Resource (TAIR): gene structure and function annotation. Nucleic Acids Res **36**, D1009–D1014.

SWINDELL, W. R., HUEBNER, M., AND WEBER, A. P. (2007). Transcriptional profiling of Arabidopsis heat shock proteins and transcription factors reveals extensive overlap between heat and non-heat stress response pathways. BMC Genomics **8**, 125.

TOUFIGHI, K., BRADY, S. M., AUSTIN, R., LY, E., AND PROVART, N. J. (2005). The Botany Array Resource: e-northerns, expression angling, and promoter analyses. Plant J **43**, 153-163.

WAHID, A., GELANI, S., ASHRAF, M., AND FOOLAD, M. R. (2007). Heat tolerance in plants: An overview. Environ Exp Bot **61**, 199-223.

WARDE-FARLEY, D., DONALDSON, S. L., COMES, O., ZUBERI, K., BADRAWI, R., CHAO, P., ET AL. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res **38**, W214–W220.

WEIGEL, D., AND GLAZEBROOK, J. (2002). *Arabidopsis: a Laboratory Manual*.(Cold Spring Harbor Laboratory Press, Cold Spring Harbor

NY).

WINTER, D., VINEGAR, B., NAHAL, H., AMMAR, R., WILSON, G. V., AND PROVART, N. J. (2007). An "electronic fluorescent pictograph" browser for exploring and analysing large-scale biological data sets. PLoS ONE **2**, e718.

WULLSCHLEGER, S. D., AND DIFAZIO, S. P. (2003). Emerging use of gene expression microarrays in plant physiology. Comp Funct Genomics **4**, 216-224.

YILMAZ, A., MEJIA-GUERRA, M. K., KURZ, K., LIANG, X., WELCH, L., AND GROTEWOLD, E. (2011). AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. Nucleic Acids Res **39**, D1118–D1122.

*«Identification of novel, functionally unresolved genes that are constitutively and specifically involved in the heat stress response, operating at the regulatory/cell nuclear level»*

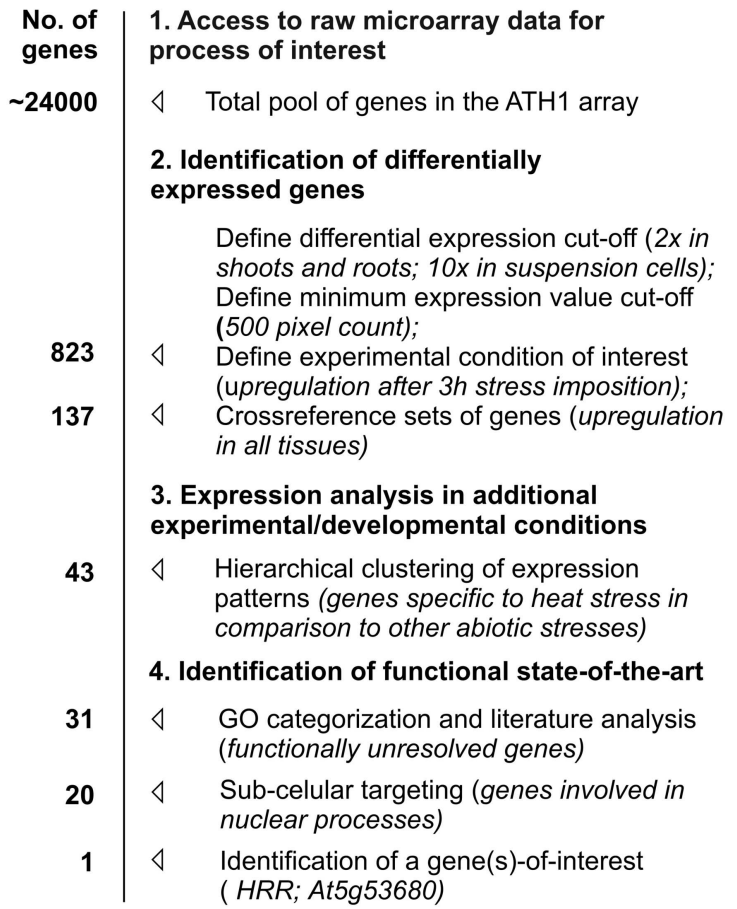


FIG. 1. Schematic outline of the strategy used in the identification of new abiotic stress determinants from publicly-available transcriptomic data.
79x122mm (600 x 600 DPI)

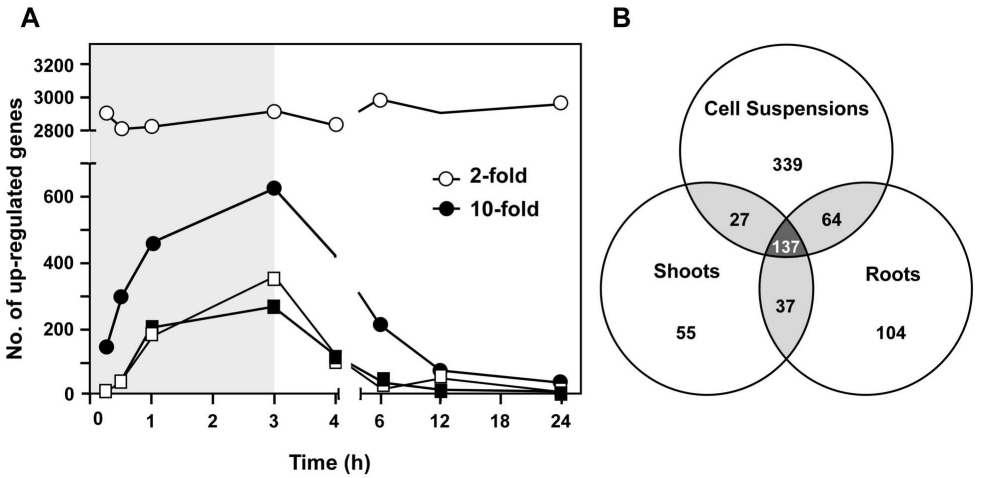


FIG. 2. Differential expression analysis after heat stress imposition in Arabidopsis thaliana. (A) Number of up-regulated genes in roots (open squares), shoots (closed squares) and cell suspensions (open and closed circles). Plants were grown for 16 days and subsequently subjected to a 3-hour heat stress treatment at 38ºC, followed by a 21-hour recovery period at 25ºC; suspension cells were grown for 6 days and subsequently subjected to a 38°C heat shock for 3 h, followed by a 21-hour recovery period at 25ºC. Cut-off values were set at >500 pixel count and up-regulated expression was defined by expression ratios above 2x (all tissues; squares and open circles) and 10x (suspension cells; closed circles). (B) Venn diagram representation of up-regulated genes within 3 h of heat stress treatment on Arabidopsis shoots, roots and cell suspensions.
84x40mm (600 x 600 DPI)

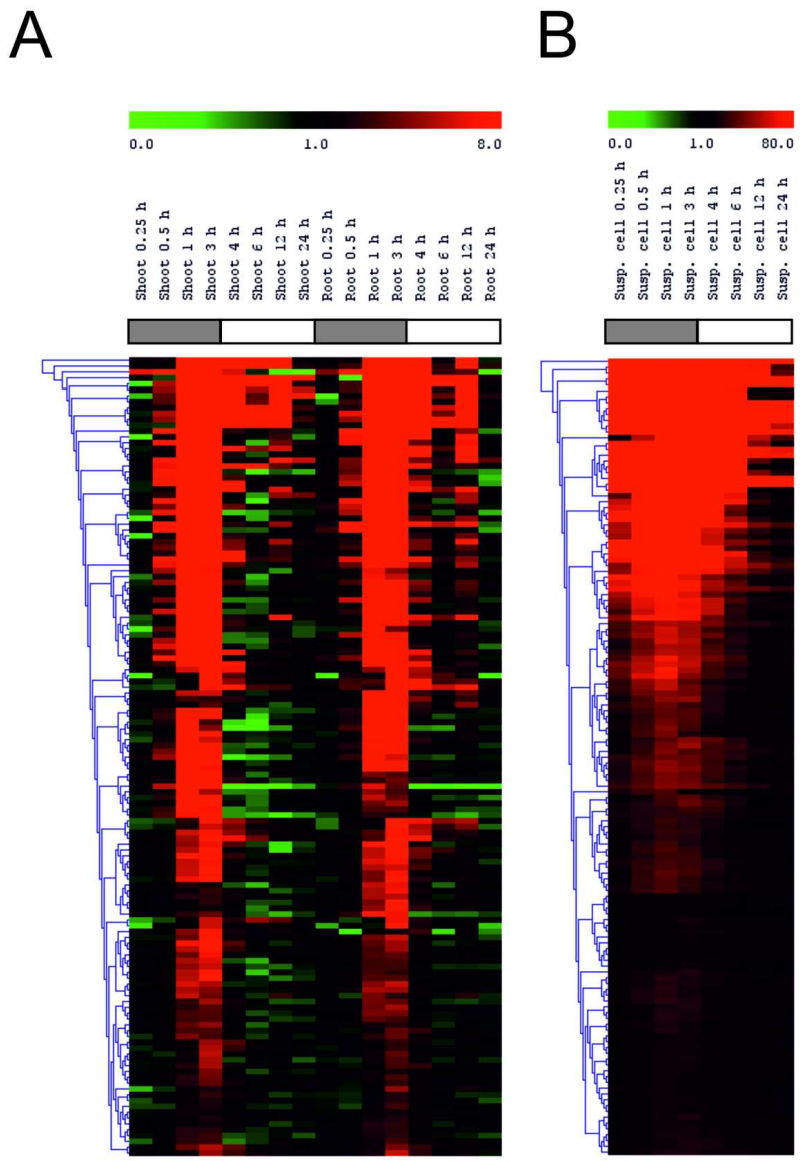


FIG. 3. Hierarchical clustering analysis of the gene expression pattern of selected 137 genes, during the time course of heat stress imposition in Arabidopsis shoots and roots (A) and suspension cells (B). Using MultiExperiment Viewer, cluster analysis was performed on gene expression ratios of heat-stressed tissues in comparison to the control (up-regulated genes are displayed in red and down-regulated genes in green). Heat stress imposition periods are highlighted in grey, recovery periods in white.
105x147mm (300 x 300 DPI)

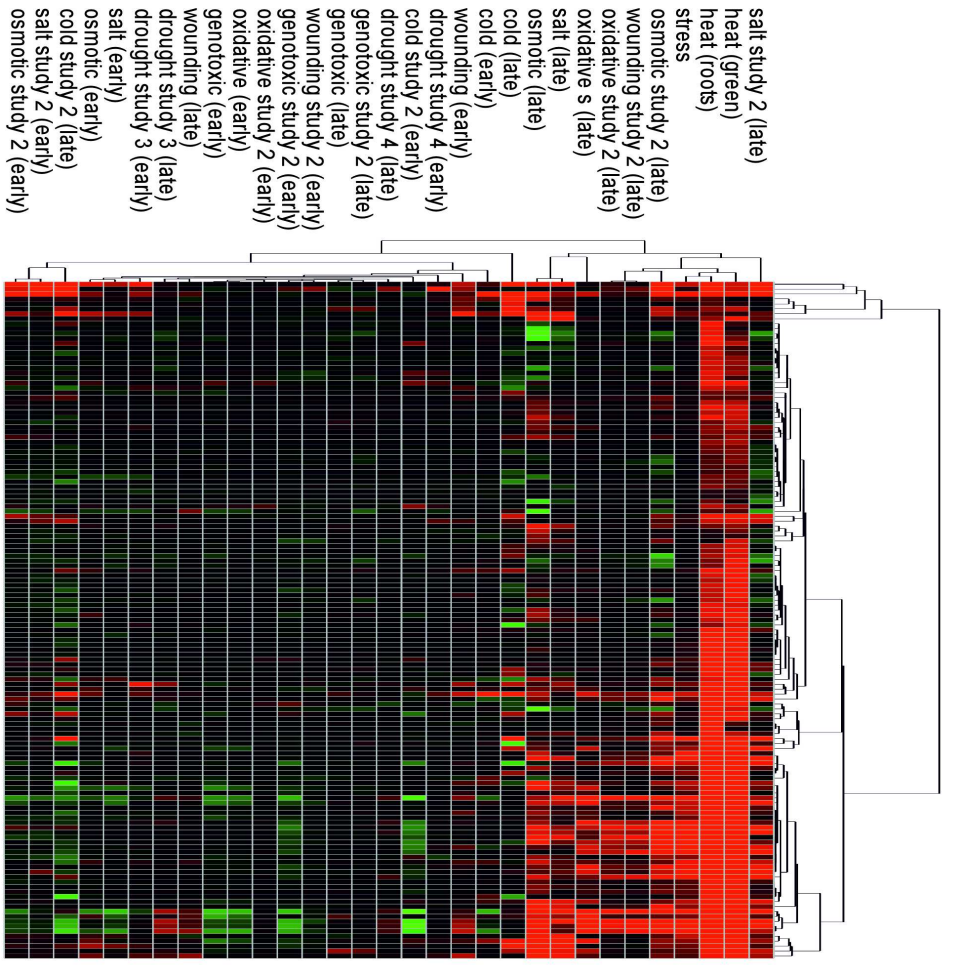


FIG. 4. Expression profile of selected 137 Arabidopsis genes in response to different abiotic stress factors, with Clustering Analysis of gene expression patterns (Genevestigator). Results are represented as heat maps for red (up-regulated) and green (down-regulated) genes.

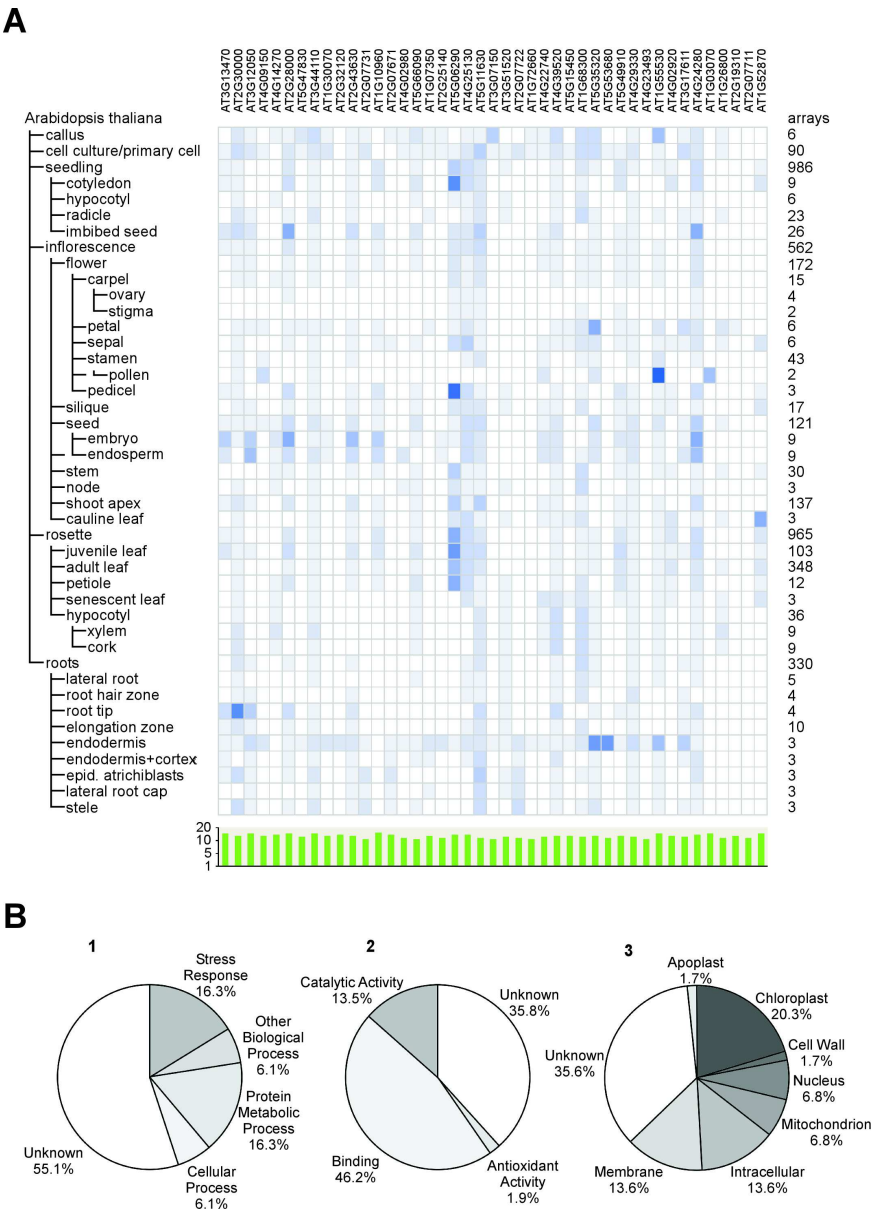


FIG. 5. Functional analysis of selected 43 heat-specific Arabidopsis genes. (A) Expression profile in different plant organs and tissues, according to the Meta-Profile Analysis (Anatomy tool) of Genevestigator. Blue intensity indicates expression level. The level of variance for each gene is indicated below the heat map and the number of arrays used to calculate the mean value is displayed on the right. (B) GO categorization according to biological process (1), cellular component (2) and molecular function (3). Analysis was performed using the GO functional annotations provided by TAIR for each selected gene. Results are presented as percentage of category frequency.

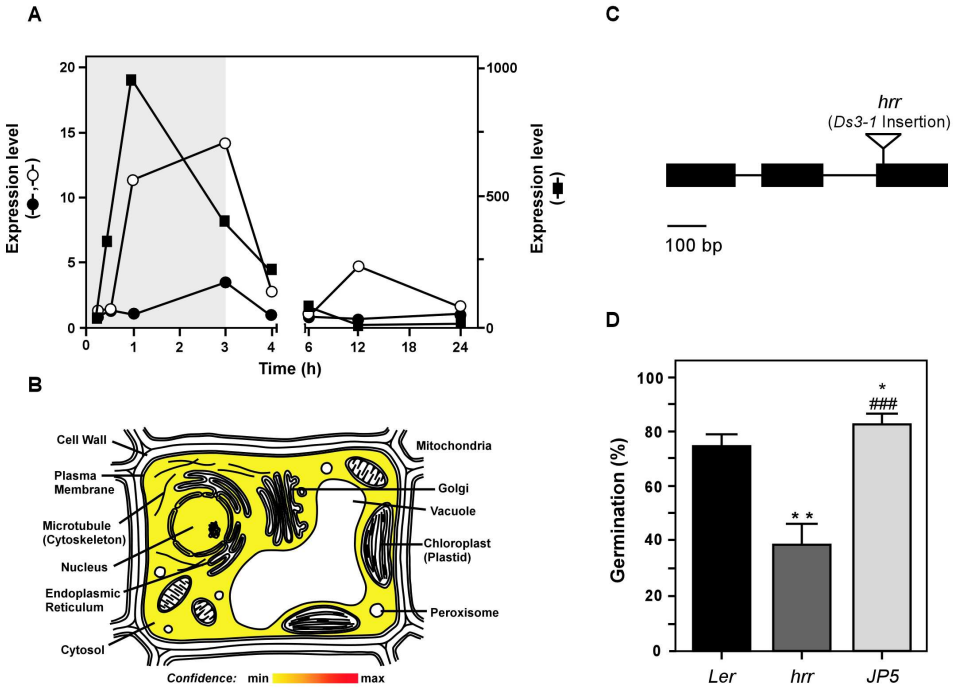


FIG. 6. Functional characterization of HRR. (A) Expression profile in Arabidopsis shoots (closed circle), roots (open circle) and cell suspensions (closed square) after a 3-hour heat stress (38°C) treatment (highlighted in grey). Expression levels are displayed as relative values (pixel count in heat stressed cells/pixel count in control cells). (B) Prediction of sub-cellular localization using Cell eFP Browser (BAR). (C) Schematic diagram depicting HRR. The hrr mutant harbours a Ds3-1 transposable element inserted in the third HRR exon. (D) The hrr knockout mutant displays heat germination impairment, while JP5 (HRR-overexpressing plants in hrr mutant background) recover seed germination capacity after heat stress imposition. Seeds were heat-stressed at 50ºC, for 60 min, and immediately sown onto MS medium. Seed germination rates were obtained by scoring the radicle emergency (4 replicates, 30 seeds each). Germinated HS-treated seeds were normalised to respective control condition (total number of germinated seeds). Data correspond to mean ± SEM. Statistics corresponds to one-way-ANOVA with a Tukey test (**, $p<0.01$; *, $p<0.5$; when compared with Ler) (###, $p<0.001$, when compared with the hrr mutant).

**Table 1.** *Arabidopsis thaliana* web-based resources for gene expression analysis towards functional discovery.

| Resource/Database (URL) | Brief description | Reference |
|---|---|---|
| ***General Resources*** | | |
| **TAIR** (www.arabidopsis.org) | The centralized resource for *Arabidopsis thaliana*. Provides access to numerous other web-based resources. Also provides access to data concerning genes, markers, maps, sequences, gene families and proteins, clones, DNA and seed stocks, polymorphisms, publications and researchers. | (Swarbreck et al. 2008) |
| **BAR** (www.bar.utoronto.ca) | Set of tools designed for expression, genomic and molecular marker analysis; specifically suitable for hypothesis generation and functional genomics research. | (Winter et al. 2007) |
| **MIPS** (mips.helmholtz-muenchen.de) | Provides annotated collections of biological data from model genomes, such as *Arabidopsis*, *Medicago*, lotus, rice, maize, tomato, barley, *Brachypodium* and *Sorghum*. This database presents several protein interaction data sets (MPCAT, MPPI and CORUM), information from micro-RNA-related phenotypes (PhenomiR), homology relations (SIMAP and PEDANT) and other versatile tools for proteomics and functional genomics (PPLIPS and CCancer), among others. | (Mewes et al. 2010) |
| **PlantGDB** (www.plantgdb.org/site/) | Develops plant species-specific EST and GSS databases, supplying web-accessible tools, facilitating inter-species queries, and also allowing genome browsing and annotation capabilities. | (Duvick et al. 2008) |
| **EnsemblPlants** (plants.ensembl.org/index.html) | Includes gene annotation, microarray probeset mapping, variation, comparative genomics (*e.g.* phylogenetic trees), and detailed information about transcripts and proteins (domains and features). | (Kersey et al. 2010) |
| ***Microarray bulk data retrieval*** | | |
| **NASCArrays** (affymetrix.arabidopsis.info/narrays/experimentbrowse.pl) | Provides access to results from Arabidopsis microarray experiments mainly run by the NASC Affymetrix Facility; the database also includes experimental data from other centers. | (Craigon et al. 2004) |
| **ArrayExpress** | Repository for transcriptomics data, available for browsing, querying and retrieval of specific | (Brazma et al. |

| | | |
|---|---|---|
| (www.ebi.ac.uk/microarray-as/ae) | experiments. | 2001) |
| **AtGenExpress** (www.weigelworld.org/resources/microarray/AtGenExpress) | Database with raw microarray data of systematic transcriptomics of Arabidopsis (abiotic stress, ecotypes, development, hormones, light and pathogen series), with descriptions of the samples and the conditions used. Data can be downloaded or visualized with the AtGenExpression Visualization Tool (AVT). | (Schmid et al. 2005; Kilian et al. 2007) |
| **GEO** (www.ncbi.nlm.nih.gov/geo/) | Functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Includes tools to help users query and download experiments and curated gene expression profiles. | (Barrett et al. 2011) |
| ***Gene expression analysis*** | | |
| **TAIR - Data Mining Tools** (www.arabidopsis.org/portals/expression/microarray/microarrayExpressionV2.jsp) | Centralized repository for expression tools, such as the TAIR Expression Search, the Arabidopsis Tiling Array Transcriptome Express Tool, the CSB.DB (Comprehensive Systems-Biology Database), GEO Profiles, Expression Angler (BAR), Expression Browser (BAR), MapMan, NASCArrays Gene Swinger, NASCArrays Two Gene Scatter Plot, NASCArrays Spot History Pathway Tools Omics Viewer, Correlated gene (ATTED-II) and Cluster Cutting among others. | --- |
| **Genevestigator** (www.genevestigator.ethz.ch) | Expression database including a suite of user-friendly tools that convert high-quality microarray data into easily interpretable results; allows the study of expression and regulation of genes in a broad variety of contexts. Has access to systematic transcriptomics data such as AtGenExpress. Though partially free, access to all tools requires an account. | (Hruz et al. 2008) |
| **BAR – Expression browser** (142.150.214.117/affydb/cgi-bin/affy_db_exprss_browser_in.cgi) | AGI codes are submitted to graphically display the gene expression profiles in numerous conditions. Cluster analysis of the data is possible. Access to systematic transcriptomics data such as AtGenExpress. | (Toufighi et al. 2005) |
| **BAR - eFP browser** (142.150.214.117/efp/cgi-bin/efpWeb.cgi) | User-friendly electronic fluorescent pictograph allows an immediate understanding of transcriptomics data. The eFP browser is presented for various developmental stages and external conditions. | (Winter et al. 2007) |
| ***Repositories of functional information*** | | |
| **BAR - Cell eFP browser** (142.150.214.117/cell_efp/cgi- | User-friendly fluorescent pictograph that gathers information from a significant number of sub-cellular localization algorithmetic predictors. Can redirect to the SUBA database. | (Winter et al. 2007) |

| | | |
|---|---|---|
| bin/cell_efp.cgi) | | |
| **BAR - Classification SuperViewer** (142.150.214.117/ntools/cgi-bin/ntools_classification_superviewer.cgi) | Creates an outline of the functional classification of a submitted list of AGI codes based on the GO database. A ranking score is also calculated for each functional class, and the input set presents over- or under-represented reliability. | (Provart and Zhu 2003) |
| **AGRIS** (arabidopsis.med.ohio-state.edu/) | Assembles information on promoter sequences, transcription factors and their target genes; three interlinked databases are available: AtTFDB, AtcisDB and AtRegNet, which give understandable and updated information on transcription factors, predicted and experimentally verified cis-regulatory elements and their interactions, respectively. | (Yilmaz et al. 2011) |
| **GeneMANIA** (genemania.org/) | Identifies related genes to a set of input genes. Uses functional association data such as protein and genetic interactions, pathways, co-expression, co-localization and protein domain similarity. | (Warde-Farley et al. 2010) |
| ***Resources for reverse genetics*** | | |
| **SIGnAL T - DNA Express** (signal.salk.edu/cgi-bin/tdnaexpress) | Platform of choice for searching the Arabidopsis genome for the localization of mutants from a large collection of mutant lines, as well as identify available cDNA sequences, using a simple map-based graphical interface. | |
| **SIGnAL T - DNA Primer Design** (signal.salk.edu/tdnaprimers.2.html) | PCR primer designing tool for efficient screening of homozygous lines in insertion mutants; outcomes include insertion site location, primer sequence and estimated product size. | |
| **TAIR - Sequence viewer** (www.arabidopsis.org/servlets/sv) | Tool for viewing the annotation of the Arabidopsis genome at the nucleotide level and retrieve nucleotide sequences. It also presents annotation units, transcripts, polymorphisms, T-DNA/Transposon insertions, and markers on the Arabidopsis genome sequence. | (Rhee et al. 2003) |
| **TAIR – Gbrowse** (gbrowse.arabidopsis.org/cgi-bin/gbrowse/arabidopsis/) | Advanced tool for browsing the Arabidopsis genome. Features include gene annotation, genomic features, endogenous transposable elements, expression data (*e.g.* SAGE), methylation and phosphorylation, orthologs and gene families, sequence similarities, variation in ecotypes. Includes the *A. thaliana* mitochondrial and chloroplast genomes. Recently the tool was extended to other plant species (*Arabidopsis lyrata*, *Brachypodium distachyon*, *Oryza sativa japonica*, *Oryza sativa indica*, *Populus trichocarpa*, *Physcomitrella patens*, *Sorghum bicolor*, *Vitis vinifera*, *Zea mays*). | (Swarbreck et al. 2008) |
| **AGRIKOLA** (www.agrikola.org) | Provides information about high-throughput cloning of Arabidopsis Gene Sequence Tags into RNAi gene silencing vectors. RNAi constructs are used to transform Arabidopsis and images of | (Hilson et al. 2004) |

| | | |
|---|---|---|
| | silenced lines can be queried by AGI code, CATMA code, gene name, or gene function. | |
| **Web MicroRNA Designer (WMD)** (wmd3.weigelworld.org/cgi-bin/webapp.cgi) | Web-based tool for artificial microRNA (amiRNA) design, for subsequent use in gene inactivation. | (Schwab et al. 2006) |
| ***Seed stock centers*** | | |
| **NASC** (arabidopsis.info/) | Collects and distributes seeds to Europe along with additional biological material and information resources, in a coordinated activity with ABRC. | (Scholl et al. 2000) |
| **ABRC** (abrc.osu.edu/) | Collects and distributes seeds, DNA clones and libraries to North America; the ordering is made in coordination with TAIR. | (Scholl et al. 2000) |
| **FLAG** (dbsgap.versailles.inra.fr/portail/) | Collects and distributes a set of FLAG mutants, more than 500 ecotypes and several recombinant inbred line populations created in the Versailles resource center (INRA). | (Samson et al. 2002) |
| **GABI-Kat** (www.gabi-kat.de/) | Distributes confirmed T-DNA insertion mutants; these lines are in process of donation to NASC. | (Li et al. 2007) |

Deleted:

**Supplementary Table S1.** Functional assignment of the selected 43 Arabidopsis heat-specific genes, with the identification of 31 functionally unresolved genes, as well as the indication of predicted sub-cellular targeting. Selected nuclear genes (20) are highlighted in solid green. HRR is highlighted in orange.

| AGI ID | Affymetrix ID | Annotation | Functionally unresolved (1) | Sub-cellular targeting prediction (2) |
|---|---|---|---|---|
| **At1g03070** | 263164_at | Bax inhibitor-1 family protein | yes | PLM, EC |
| **At1g07350** | 261081_at | RNA-binding (RRM/RBD/RNP motifs) family protein | yes | Mit, Nuc, Chl |
| **At1g10960** | 260481_at | ATFD1_FD1__ferredoxin 1 | | Mit, Nuc, Chl |
| **At1g26800** | 261265_at | RING/U-box superfamily protein | yes | Nuc |
| **At1g30070** | 260025_at | SGS domain-containing protein | yes | Nuc, PLM |
| **At1g52870** | 260155_at | Peroxisomal membrane 22 kDa (Mpv17/PMP22) family protein | yes | Cyt, Mit, Px, Chl |
| **At1g55530** | 265077_at | RING/U-box superfamily protein | yes | Nuc |
| **At1g68300** | 260444_at | Adenine nucleotide alpha hydrolases-like superfamily protein | yes | Cyt, Px, Chl |
| **At1g72660** | 259913_at | P-loop containing nucleoside triphosphate hydrolases superfamily protein | yes | Cyt, Mit |
| **At2g07671** | 257339_s_at | ATP synthase subunit C family protein | yes | Ec, Glg, Mit, Nuc |
| **At2g07711** | 257338_s_at | pseudogene, similar to NADH dehydrogenase subunit 5 | yes | |
| **At2g07722** | 266014_s_at | unknown protein | yes | Ec, Mit, Nuc |
| **At2g07731** | 244953_s_at | pseudogene, similar to NADH-ubiquinone oxidoreductase chain 6 | yes | |
| **At2g19310** | 267336_at | HSP20-like chaperones superfamily protein | | Cyt, Mit, Nuc, Chl |
| **At2g25140** | 264402_at | CLPB-M_CLPB4_HSP98.7__casein lytic proteinase B4 | | Mit, Chl |
| **At2g28000** | 264069_at | CH-CPN60A_CPN60A_SLP__chaperonin-60alpha | | Cyt, Mit, Chl, Vac |
| **At2g30000** | 266806_at | PHF5-like protein | yes | Cyt, Ec |
| **At2g32120** | 265675_at | HSP70T-2__heat-shock protein 70T-2 | | Cyt, Chl |
| **At2g43630** | 260586_at | unknown protein | yes | Cyt, Mit, Nuc, Chl |
| **At3g07150** | 258827_at | unknown protein | yes | Mit, Nuc |
| **At3g12050** | 256663_at | Aha1 domain-containing protein | | Cyt, Mit, Px, Chl |

| | | | | |
|---|---|---|---|---|
| **At3g13470** | 256983_at | TCP-1/cpn60 chaperonin family protein | yes | Mit, Nuc, Chl |
| **At3g17611** | 258406_at | ATRBL14_RBL14__RHOMBOID-like protein 14 | yes | ER, Ec, Mit, Chl |
| **At3g44110** | 252670_at | ATJ_ATJ3__DNAJ homologue 3 | | Mit, Nuc, PLM |
| **At3g51520** | 252064_at | diacylglycerol acyltransferase family | yes | Mit, Chl, PLM |
| **At4g02920** | 255456_at | unknown protein | yes | Cyt, Mit, Nuc |
| **At4g02980** | 255412_at | ABP_ABP1__endoplasmic reticulum auxin binding protein 1 | | ER, Ec, Chl |
| **At4g09150** | 255077_at | T-complex protein 11 | yes | Cyt, Nuc |
| **At4g14270** | 245602_at | Protein containing PAM2 motif which mediates interaction with the PABC domain of polyadenyl binding proteins. | yes | Cyt, Nuc |
| **At4g22740** | 254278_at | glycine-rich protein | yes | Nuc, Chl, PLM |
| **At4g23493** | 254263_at | unknown protein | yes | Cyt, Mit, Nuc, Chl |
| **At4g24280** | 254148_at | cpHsc70-1__chloroplast heat shock protein 70-1 | | Mit, Nuc, PLM, Chl |
| **At4g25130** | 254099_at | PMSR4__peptide met sulfoxide reductase 4 | yes | Mit, Nuc,chl |
| **At4g29330** | 253712_at | DER1__DERLIN-1 | yes | Ec, Chl |
| **At4g39520** | 252883_at | GTP-binding protein-related | yes | Cyt, Mit, Px |
| **At5g06290** | 250733_at | 2-Cys Prx B_2CPB__2-cysteine peroxiredoxin B | | Ec, Mit, Chl |
| **At5g11680** | 250332_at | unknown protein | yes | Cyt, Mit, Nuc, PLM |
| **At5g15450** | 246554_at | APG6_CLPB-P_CLPB3__casein lytic proteinase B3 | | Cyt, Mit, Chl |
| **At5g35320** | 246612_at | unknown protein | yes | Nuc |
| **At5g47830** | 248774_at | unknown protein | yes | Nuc, Px |
| **At5g49910** | 248582_at | cpHsc70-2_CPHSC70-2EAT SHOCK PROTEIN 70-2_HSC70-7__chloroplast heat shock protein 70-2 | | Mit, Nuc, Chl |
| **At5g53680** | 248215_at | RNA-binding (RRM/RBD/RNP motifs) family protein | yes | Cyt, Nuc |
| **At5g66090** | 247139_at | unknown protein | yes | Mit, Nuc, Chl |

(1) Genes without information on gene/protein function and not previously associated with heat stress responses.

(2) Chl - chloroplast; Mit - mitochondria; Nuc - nucleus; PLM- plasma membrane; Ec- extracellular; Px- Peroxissome; Cyt- cytosol; Glg- Golgi; Vac- Vacuole; ER- Endoplasmic reticulum (according to Cell eFP Browser, BAR).